Rényi Alfréd

Információelmélet és nyelvészet

1. Először arról szeretnék néhány szót szólni, hogy mi is az információelmélet. Lehet beszélni információelméletről szűkebb és tágabb értelemben. A szűkebb értelemben vett információelmélet a matematikának, közelebbről a valószínűségszámításnak egy új fejezete, amely az utolsó tizenöt évben jött létre és amelynek létrehozása Claude Shannon[1] nevéhez fűződik. A matematikának ez a fejezete a híradástechnikában előforduló információ-átviteli folyamatok matematikai elméleteként alakult ki; az információ-továbbítás, információ-tárolás, információfeldolgozás problémái alkotják tárgyát.

A szűkebb értelemben vett információelmélet tehát az üzenetközvetítéssel foglalkozik (függetlenül attól, hogy az üzenetközvetítés ember és ember között, vagy ember és gép, vagy gép és gép között történik) és azt vizsgálja, hogyan lehet az információt a legcélszerűbb alakban bizonyos zavaró külső tényezők („zajok") jelenlétében közvetíteni, hogy kell ehhez „kódolni", hogy kell azután a kapott üzenetet „dekódolni", stb. Ezt a szűkebb értelemben vett információelméletet a maga területén (a híradástechnikában, az elektronikus számológépek, az automatizálás elméletében stb.) nagy sikerrel alkalmazzák.

Ez a speciálisabb értelemben vett információelmélet is sok tekintetben nyújthat segítséget a nyelvtudománynak, azonban a nyelvésznek tulajdonképpen többre van szüksége: olyan matematikai elméletre, amely a nyelv mélyebb összefüggéseit is képes megragadni. Azt az elméletet, amely az információt nemcsak kvantitatív szempontból, hanem tartalmilag is vizsgálja, nevezhetjük tágabb értelemben vett információelméletnek. Már itt hangsúlyoznom kell azonban, hogy míg a szűkebb értelemben vett információelmélet egy kidolgozott matematikai diszciplína, amelyben rengeteg nyitott probléma van ugyan, de azért már ma is egy kerek és ha nem is lezárt, de fejlett elmélet, ezzel szemben az, amit tágabb értelemben vett információelméletnek nevezek, csak az első lépéseknél tart. Ez a tágabb értelemben vett információelmélet lényegében úgy jött létre, hogy észrevették, hogy a szűkebb értelemben vett információelméleti

[1] C. E. Shannon—W. Weaver: The mathematical theory of communication (Urbana, Ill., 1949).

eredményeknek a hordereje még nagyobb, mint amire a kiindulását képező problémákból következtetni lehetne. Ezzel persze korántsem akarom lebecsülni azoknak a műszaki problémáknak a jelentőségét, amelyek megoldására az információelmélet létrejött, hiszen ezek a maguk területén nagyon fontos problémák. A matematikában azonban nem meglepő, hogy egy területen kidolgozott módszerekről később kiderül, hogy egész távoleső területeken is felhasználhatók. Ugyanakkor persze ki szokott derülni az is, hogy az új területen való felhasználás érdekében e módszerek általánosítására, kiegészítésére, újabb kutatásokkal való továbbfejlesztésére van szükség. Ez a helyzet ma az információelmélet terén is. A tágabb értelemben vett információelmélet tárgyát képezheti a nyelv nemcsak mint szimbólumok sorozata (mert mint ilyen, már a szűkebb értelemben vett információelméletnek is tárgyát képezi), hanem mint tartalommal, jelentéssel bíró jelsorozat. A nyelv statisztikai tulajdonságai, statisztikai törvényei önmagukban is fontosak és érdekesek, és feltétlenül az eddiginél több figyelmet érdemelnek. Azonban a matematikai nyelvészetben ma ennél még többre törekszenek, tudniillik, hogy valamilyen módon a nyelv tartalmát is matematikailag tárgyalják.

Shannon nagyon nyomatékosan hangsúlyozta, hogy amikor egy mértékszámot adott meg, amivel az információ mennyiségét mérjük, akkor az információnak pusztán a mennyiségét vizsgálta, teljesen eltekintve a tartalmától, még inkább az értékétől. Az általánosabb értelemben vett információelméletnek tárgyát képezheti nemcsak a nyelv, hanem egy egész sereg más emberi tevékenység is, például az, ahogy egy orvos diagnózist állít fel egy betegről, a különböző vizsgálatokból nyert információkat összegezi, összehasonlítja, feldolgozza. Tárgyát képezheti ennek a tágabb értelemben vett információelméletnek például az a folyamat, ahogy egy vizsgálóbíró nyomoz. Tárgyát képezi továbbá az átöröklés, mint (genetikai) információátadási folyamat; ez egyik viszonylag eléggé kialakult fejezete ennek az általános információelméletnek. Tárgyát képezhetik a dokumentáció gépi módszerei, továbbá a tanulás folyamata, beleértve nemcsak az ember, hanem a gép „tanulását" is. Ez a tágabb értelemben vett információelmélet kétségtelenül egy rendkívül érdekes és izgalmas téma, de — még egyszer hangsúlyozom — sokkal kevésbé kialakult állapotban van ma, mint a szűkebb értelemben vett információelmélet.

Az előadásban először a szűkebb értelemben vett információelméletről szeretnék egész röviden beszélni, azután néhány szót szólni a matematikai nyelvészet perspektíváiról. Mielőtt erre rátérnék, szeretném azt hangsúlyozni, hogy a tágabb értelemben vett információelmélet csak a szűkebb értelemben vett információelméletre épülhet, hiszen nem indulhatunk ki másból, mint ami van, és csak arra próbálhatunk valami újat építeni. A kétfajta információelmélet közötti különbséget illetőleg döntőnek az információ tartalmának a kérdését érzem: amíg a tartalommal nem foglalkoztunk, vizsgálatunk a szűkebb értelemben vett információelméletbe tartozik, de ha a tartalommal is foglalkozunk, az már a tágabb értelemben vett információelméletbe tartozik. Egy hasonlattal lehet legjobban megvilágítani a helyzetet; ha valaki elmegy a postára táviratot feladni, akkor a postatisztviselő megszámolja a szavak számát, kirója a díjat és utána

leadja a táviratot Morse-jelekkel, igyekszik ezt hibátlanul továbbítani; azzal, hogy a táviratban mi van, ő nem foglalkozik (és nem is volna helyes, ha foglalkozna). Ma már technikailag az is megvalósítható, hogy az egész folyamat gépesítve legyen, és a távirat úgy érkezzék meg a címzetthez, hogy azt senki más ne is lássa; így ma már lehetséges, hogy a levéltitok mintájára a „távirat-titok" is létrejöjjön, és ez kívánatos is volna. A távirattovábbítással kapcsolatos minden probléma a szűkebb értelemben vett információelméletbe tartozik. Általános információelméleti jellegű tevékenység volna azonban a távirat-feladással kapcsolatban az olyan postatisztviselő munkája, akinek az volna a feladata, hogy tanácsot ad az embereknek, hogy hogyan lehet a táviratot más szavakkal, rövidebben megfogalmazni. Elvileg elképzelhető az is, hogy ezt a feladatot egy gép végezze el. Kétségtelen, hogy az információ tartalmának kérdése sokkal bonyolultabb, komplexebb dolog, mint a puszta mennyiségének a kérdése. Így például az információmennyiség teljesen objektív fogalom, míg a tartalom és a jelentés kérdésénél sokkal nehezebb a szubjektív és objektív elemeket elválasztani.

2. Ezen bevezetés után most a szűkebb értelemben vett információelméletről szeretnék néhány szót szólni. Ennek az elméletnek a kiindulópontja az a kérdés volt, hogy hogyan lehet az információ mennyiségét számszerűleg kifejezni, tehát egy mérőszámot adni arra, hogy egy közlés mennyi információt tartalmaz. Ebből a célból először egy egységre van szükség: az információ egységére. Ilyen egység, mint közismert, az úgynevezett „bit", amelyen az az információ értendő, amely egyetlen olyan jellel kifejezhető, amelynek két lehetséges értéke van, a 0 vagy az 1. Más szavakkal: az információ egysége egy olyan válaszban foglalt információ mennyisége, amely válasz *igen* vagy *nem* szóval kifejezhető, teljesen függetlenül a kérdés tartalmától. Egy tetszőleges közlésben foglalt információ mennyiségét számszerűleg úgy lehet mérni, hogy azt a közlést kifejezzük ilyen két értékű (0 vagy 1 értéket felvenni képes) jelek sorozatával és a jelek számával mérjük az információ mennyiségét. Ez minden fajta közlés esetében elvileg is és gyakorlatilag is lehetséges. Amennyiben számokról van szó, akkor át lehet írni a számokat a kettes számrendszerbe. Ha az információ írott szöveg, meg lehet számozni a betűket és írásjeleket (a szóközt is beleértve), és ki lehet fejezni a betűk (írásjelek) sorszámát a kettes számrendszerben. Persze minden közlés sokféleképpen fejezhető ki 0 és 1 jelek sorozatával; egy közlésben foglalt információ mennyiségén azt értjük, hogy a közlést a leggazdaságosabb módon kifejezve 0 és 1 jelekkel, hány jelből fog állni. Könnyen be lehet látni, hogy ha egy „nyelv"-ben n jel van (n „betűből" álló ábécéje van a nyelvnek), akkor e „nyelv" egy betűjének átírásához átlagosan $\log_2 n$ 0- vagy 1-jel szükséges ($\log_2 n$ az n szám 2 alapú logaritmusát jelöli). Az $I = \log_2 n$ formulát nevezik Hartley-féle formulának. A Hartley-formula csak abban az esetben alkalmazható, ha a szóban forgó nyelv „betűi" egyformán valószínűek, azaz általában mind körülbelül ugyanolyan gyakran fordulnak elő. Az általános esetben, amikor a „nyelv" betűinek valószínűségei a $P_1, P_2, \ldots, P_n$ számok, a nyelv egy betűjének átírásához átlagban

$$I = P_1 \log_2 \frac{1}{P_1} + P_2 \log_2 \frac{1}{P_2} + \ldots + P_n \log_2 \frac{1}{P_n}$$

0- vagy 1-jelre van szükség; ez az úgynevezett Shannon-féle formula. Ezzel még korántsem merült ki az információelméletben használt információ-mértékszámoknak az összessége; a Shannon-félétől különböző mértékszámok is vannak,[2] melyek néha hasznosak ugyanezen információnak a mérésére. De vannak más fogalmak is, amelyekre itt nem térhetek ki részletesen, mint például a feltételes információ, az információnyereség és a relatív információ fogalma.

Az információelmélet legalaposabban kidolgozott ága a zajos csatornán keresztül való információ-továbbítás elmélete. A következő sémával szokták jellemezni ezt a feladatot: adva van egy információ-forrás, amely bizonyos jeleket (szimbolumokat) produkál, ezeket kell továbbítani egy „zajos" csatornán keresztül. Azon, hogy a csatorna „zajos", az értendő, hogy a továbbítás folyamán a jelek eltorzulhatnak, és a felvett jel nem lesz azonos a leadott jellel. Továbbá egy ilyen csatornának megvan az a jellegzetessége, hogy csak meghatározott típusú jelek közvetítésére képes, és csak meghatározott típusú jeleket szolgáltat; szóval egy ilyen csatornának van egy bemenő ábécéje és van egy kimenő ábécéje. Így az információ-forrás által produkált jeleket valamilyen módon át kell alakítani olyan jelekké, amelyeket a csatorna képes továbbítani; ez a kódolás. Ezt oly módon kell elvégezni, hogy ez az átalakítás a zajokkal szemben minél „ellenállóbb" legyen. Azután a csatorna másik végénél a felvett jeleket dekódolni kell, tehát át kell alakítani őket olyan jelekké, amilyeneket a felvevő kíván. Például a televíziónál az eredeti jel egy kép, ezt a képet felbontja a felvevőgép először is kis elemekre, ezeket az elemeket azután átalakítja elektromágneses hullámokká, ezek az elektromágneses hullámok terjednek a téren át, ami ez esetben a „csatorna" szerepét tölti be; a légköri zavarok, a légköri elektromosság, egyéb adók működése következtében ezek a jelek kissé eltorzulva érkeznek meg a televíziós készülékhez, amely visszaalakítja őket képpé.

Az információelméletben szerepel egy nagyon fontos fogalom, amiről még szólni szeretnék, ez a r e d u n d a n c i a fogalma. Az előbb azt hangsúlyoztam, hogy egy közleményben foglalt információ mennyiségének a megállapításánál azt nem akárhogy kell átírni 0 és 1 értékű jelekből álló sorozattá, hanem a leggazdaságosabb módon. Mármost arról, hogy egy szöveg az adott jelkészlet segítségével a leggazdaságosabb módon van-e kifejezve, nemcsak 0 és 1 jelekből álló szöveg, hanem tetszőleges

[2] Vö. Rényi Alfréd, Az információelmélet néhány alapvető kérdése: A Magyar Tudományos Akadémia Matematikai és Fizikai Osztályának Közleményei 10. 251—282 (1960); On measures of entropy and information: Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability (University of California Press, 1961), I, 547—61; Az információ-akkumuláció statisztikus törvényszerűségeiről: A Magyar Tudományos Akadémia Matematikai és Fizikai Osztályának Közleményei 12. 15—31 (1962); Egy információelméleti problémáról: A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei 6 B. 505—16 (1961).

szimbólumokból álló szöveg esetében is lehet beszélni. Ha a szimbólumok felhasználása nem a leggazdaságosabban történik, akkor azt mondjuk, hogy a szöveg redundáns; mégpedig a redundancia számszerűleg azt adja meg, hogy a szövegnek tulajdonképpen hányadrésze az, ami felesleges, tehát hányadrészével válna rövidebbé a szöveg, ha a leggazdaságosabban használnánk fel a rendelkezésünkre álló jelkészletet.

Ebben az értelemben a természetes nyelvek persze erősen redundánsak; ezt azonban semmiképpen sem szabad a nyelv rovására írni. Éppen ellenkezőleg, a redundanciának a nyelvben rendkívül pozitív szerepe van, és egy redundancia nélküli nyelv gyakorlatilag használhatatlan volna. A redundanciára tehát szükség van, például azért, mert ha egy mesterséges nyelv abszolút redundanciamentes volna, az azt jelentené, hogy nemcsak az igaz benne, hogy minden szimbólum egyformán valószínű, hanem az is igaz, hogy egyes szimbólumok egymástól teljesen függetlenül követhetik egymást. Egy ilyen „nyelv"-ben tehát minden elképzelhető jelsorozat értelmes szöveg. Viszont nyilvánvaló, hogy a betű- (illetőleg hang- vagy fonéma-)sorozatok túlnyomó része kiejthetetlen. Tehát már ezért sem képzelhető el egy redundancia nélküli nyelv. Másodszor, ha egy „nyelv"-ben a betűk tetszőlegesen választott sorozatának értelme volna, ezáltal nagyon megnőne a félreértés lehetősége. Azáltal, hogy a természetes nyelv redundáns, elég nagy százalék sajtóhibát szinte teljes biztonsággal korrigálni lehet. Szóval, bár az információelméletben általában célkitűzés a felesleges redundancia megszüntetése, ez nem lehet olyan célkitűzés, amit a természetes nyelveknek követnie kellene.

A természetes nyelv redundanciája teszi lehetővé, hogy megértsünk egy olyan szöveget, amelynek — azáltal, hogy esetleg a helyiségben zaj van — csak egy elég kis százalékát halljuk ténylegesen, talán csak 50—60 százalékát. Ha tisztában vagyunk is azzal, hogy a redundancia a természetes nyelvben szükséges és pozitív funkciói vannak, mindenesetre érdemes azt tudni, hogy ez a redundancia egy szövegben mekkora. Arra vonatkozólag, hogy ezt hogyan lehet meghatározni, a Magyar Nyelvtudományi Társaságban tartott előadásomban[3] ismertettem egy eljárást. Hangsúlyozni kell ezzel kapcsolatban, hogy a redundancia függ a szövegtípustól. Lehet beszélni általában a magyar nyelv redundanciájáról, de ezen belül lehet külön a szépirodalomnak vagy a napi sajtónak a redundanciájáról, a szépirodalmon belül lehet az egyes írók stílusának a redundanciájáról beszélni, és ezek a számértékek egymástól elég nagy mértékben eltérhetnek. A redundancia-meghatározás is felhasználható az egyes stílusok közötti különbségek kimutatására. A redundancia kérdését tulajdonképpen mint tipikus példát említettem arra, hogy milyen eredményeket adhat a szűkebb értelemben vett információelmélet fogalmainak alkalmazása a természetes nyelvekre. Kétségtelenül ez is vezet kézzelfogható, nyelvtudományi, sőt esetleg még irodalmi szempontból is érdekes eredményekre.

[3] Információelmélet és nyelvtudomány. Előadás a Magyar Nyelvtudományi Társaság általános nyelvészeti és fonetikai szakosztályában 1962. február 20-án. Vö. Magyar Nyelv 59. 121 (1962).

Szeretném azt is hangsúlyozni, hogy a redundanciába a nyelvtani struktúra nagyon erősen belejátszik, szóval a redundancia nemcsak egyszerűen a betű-, illetőleg szóstatisztikától függ. Hiszen ha egy bizonyos szónak néhány betűje meg van adva, akkor gyakran a nyelvtani szabályok alapján ki lehet találni a hiányzó betűket. Tehát a redundancia fogalmára nem lehet azt mondani, hogy ez csak a nyelv felszínes jelenségeit ragadja meg, mint mondjuk a betű-gyakoriságok. De azért ez a fogalom még mindig nem hatol túlságosan mélyre.

Mint mondottam, az általános információelmélet még csak a kezdeti lépéseket tette meg; gondolok itt például Bar-Hillel munkáira. Még távol vagyunk attól például, hogy olyan kérdéseket, mint az irodalmi nyelvben az utalásokkal való információközlést, matematikailag tárgyalni tudnánk. Egy szó tartalmához ugyanis hozzátartoznak azok az összefüggések, amelyekben az illető szót a mindennapi életben használják, azok az összefüggések, amelyekben az illető szó (különösen egy-egy ritka szó) valamilyen közismert klasszikus irodalmi műben szerepel, és ezek segítségével az író fel tudja idézni az egyes szavakhoz az irodalmi köztudatban fűződő hangulatot, atmoszférát. Különösen gyakran élnek ezzel az eszközzel — amelyet rezonancia-keltésnek is nevezhetünk — a költők. Az információ-közlésnek sok ilyen rejtettebb módja van, amelyet szintén nem szabad figyelmen kívül hagyni a nyelv információelméleti vizsgálatánál.

3. Befejezésül csak egész röviden még arról szeretnék beszélni, hogy mit várhatunk az információelmélet nyelvészeti alkalmazásaitól. Először is szeretném még egyszer hangsúlyozni, hogy annak, amit szűkebb értelemben vett információelméletnek neveztem, az alkalmazása (vagyis lényegében a nyelv statisztikai vizsgálata) önmagában is rendkívül értékes felvilágosításokat adhat, és nemcsak hasznos, hanem egyenesen nélkülözhetetlen például a gépi fordításban, a szótárkészítésben, de a stílusvizsgálatoknál is. Azonban van ennél egy talán még jelentősebb perspektívája az információelmélet nyelvészeti alkalmazásainak, mégpedig az, hogy az információelmélet a természetes nyelveket beállítja egy szélesebb kategóriába, más információközlési módszerek közé, és így lehetővé teszi az összehasonlítást, az egyezések és a kontrasztok kiemelését, és ezáltal elősegíti a nyelvre vonatkozó ismereteink elmélyítését.

Melyek azok az információközlési módszerek, amelyekkel párhuzamba lehet állítani a nyelvet? Ezek közt említem először a matematikának a formanyelvét, és a matematikai logikának a formanyelvét, továbbá a különböző mesterséges nyelveket, mint például az ALGOL, vagy a Lincos.[4]

A matematikai nyelvészet eredményei felhasználhatók a nyelvtanulás racionálisabb alapon való megszervezésénél is. Arra gondolok, hogy — azt hiszem — mindenki, aki valamilyen idegen nyelvet megtanult, azt legtöbbször nem úgy tanulta meg, hogy végig követte a nyelvtanulás előírt tankönyvszerű menetét, hanem egy bizonyos fokon „szabadúszóvá" vált, elkezdett olvasni, eleinte esetleg még szótárral, később azonban már szó-

[4] H. Freudenthal: Lincos. Design of a language for cosmic intercourse (Amsterdam, 1960).

tár nélkül. Aki így tesz, tulajdonképpen azt a statisztikai törvényszerűséget használja, ha nem is mindig tudatosan, hogy ha egy könyvet olvasok egy idegen nyelven, akkor az új szavaknak a száma, amelyekkel találkozom, exponenciálisan csökken. Az új szavakkal pedig az ember többször is találkozik és az összefüggésből összehasonlítással ki lehet találni a szónak a jelentését (amennyiben a szavak nagy részét ismeri az olvasó). Tehát a nyelv redundanciáját fel lehet használni a nyelv tanulásánál is.

Meg vagyok róla győződve, hogy az elkövetkező években és évtizedekben még tucatszám fognak születni a legkülönbözőbb célokra szolgáló mesterséges nyelvek és hogy ezeknek a mesterséges nyelveknek a létrehozása révén megértünk majd bizonyos dolgokat a természetes nyelvekről, olyan dolgokat, amelyekre különben nem jöttünk volna rá; ugyanúgy, ahogy a számológépek megalkotása útján értettük meg az agyműködés bizonyos jellegzetességeit. Úgy hiszem, ilyen tekintetben is nagy perspektívái vannak a matematikai nyelvészetnek.